

RESEARCH ARTICLE

Quantifying geographic domain shift to decouple the geospatial transferability of human mobility flow generation models

Zhiyong Zhou ^{a,b}, Song Gao ^{a,*}, Qianheng Zhang ^a, Feng Zhang ^b, Zhenhong Du ^b

^aGeoDS Lab, Department of Geography, University of Wisconsin-Madison, Madison, WI, USA

^bSchool of Earth Sciences & Zhejiang Key Laboratory of Geographic Information Science, Zhejiang University, Hangzhou, China

ABSTRACT

Human mobility serves as an essential proxy for understanding social, economic, and environmental dynamics in urban systems. However, since human mobility data are often scarce due to high collection costs and privacy concerns, generating them from limited mobility observations or auxiliary data sources is needed. Geospatial transferability, which measures a model's capability in a new location or unseen region, is a critical dimension for comparing different human mobility generation models. However, few studies have studied the intrinsic characteristics of geospatial transferability. To this end, this study systematically investigates the geospatial transferability of four representative human mobility generation models using a large-scale benchmark dataset of census tract-level commuting flows across 2,265 counties in the United States. Inspired by the domain adaptation theory in machine learning, we introduce geographic domain shift to describe the intrinsic differences in geographic feature distributions and spatial structures between source and target regions, which may jointly affect model transferability. Moreover, we propose two metrics, mutual information and spatial shift, to quantify the geographic domain shift. To examine their associations with model transferability, we employ linear mixed-effects regression to analyze the associations between geographic domain shifts and transferability. Our results reveal substantial spatial heterogeneity and asymmetry in transfer performance across regions. Both information shift and spatial shift exhibit statistically significant and complementary explanatory power. This indicates that geospatial transferability depends not only on model design but also on intrinsic geographic differences. These findings provide a novel methodological framework for evaluating and improving the geospatial transferability of human mobility generation models and support more robust and fair human mobility data synthesis across diverse regions. It also offers insights on spatial transferability for GeoAI model development.

KEYWORDS

Geospatial transferability; covariate shift; human mobility flow generation; transfer learning

*A preprint draft and the final version will be available on the Annals of AAG; *Corresponding Author, Song Gao, Email: song.gao@wisc.edu

1. Introduction

Human mobility is an essential proxy for spatial-social dynamics embedded in urban and transportation systems (Pappalardo *et al.* 2023, Xu *et al.* 2016, 2023, Zhu and Ma 2026), including social segregation (Nilforoshan *et al.* 2023, Xu *et al.* 2025a), socioeconomic status (Xu *et al.* 2018, Wang *et al.* 2024a, Yabe *et al.* 2025), tourism (Xu *et al.* 2021), pandemic responses (Hou *et al.* 2021, Huang *et al.* 2022, Noi *et al.* 2022, Santana *et al.* 2023), and environmental sustainability (Zheng *et al.* 2024). However, human mobility data are often scarce and limited in geographic coverage due to the high costs associated with the deployment of localization infrastructure and the significant privacy concerns about the collection and access of location data, which constrains comprehensive analyses of social dynamics. Therefore, synthesizing or generating realistic human mobility data has become as a crucial approach for mitigating data scarcity or biases and compensating for missing or inaccessible mobility records (Luca *et al.* 2021).

Although physical and statistical models have long been used to characterize individual- and population-level mechanisms of human mobility (Gonzalez *et al.* 2008, Schlöpfer *et al.* 2021, Boucherie *et al.* 2025, Chen *et al.* 2025), data-driven machine learning approaches have demonstrated stronger predictive and generative capabilities, even when human movement observations are sparse or entirely unavailable. These models leverage auxiliary static geographic information, such as satellite imagery, demographic attributes, and points of interest (POI), to infer human mobility patterns (Xu *et al.* 2025b, Rong *et al.* 2023a, 2025, Simini *et al.* 2021). However, because training data are typically drawn from a limited set of geographic areas, the geospatial transferability of human mobility generation models has become a critical concern due to the underlying data bias and spatial heterogeneity of physical and social environments (Luca *et al.* 2021, Goodchild and Li 2021).

To address this issue, existing studies have incorporated prior knowledge of human mobility behavior, such as gravity law, visitation law, and long-tail distribution of mobility flows, into deep learning models. These approaches have been shown to improve cross-region transfer learning performance when models are evaluated on datasets from regions different from the training area (Simini *et al.* 2021, Schlöpfer *et al.* 2021, Rong *et al.* 2023b, Wang and Zhu 2025, Zhao *et al.* 2025b, Yang *et al.* 2026). To date, however, assessments of model transferability typically assume that the target regions are inherently different from the training region. While this assumption facilitates comparisons among different human mobility generation models under a fixed evaluation setting, it overlooks the degree of intrinsic geographic differences between training and testing regions. In practice, a transferred geographic area is not necessarily different from the training area, as geographic difference or dissimilarity can be shaped by spatial

proximity (Tobler 1970), socioeconomic status (Goodchild 2004), physical environments (Zhang *et al.* 2024, Yan 2024), and environmental configurations (Zhu *et al.* 2018). In the context of human mobility flows, several studies have empirically identified regional effects and spatial heterogeneity of human mobility flows (Xu 2023, Chen *et al.* 2025, Long *et al.* 2025).

By overlooking such intrinsic geographic differences, existing evaluations leave unresolved how difficult it is to transfer a human mobility generation model from one region to another. This limitation can further bias the generation of human mobility data across different geographic areas, a challenge increasingly observed in geospatial artificial intelligence (GeoAI) models more broadly (Gao *et al.* 2023, Lou *et al.* 2025, Hou *et al.* 2025, Wang *et al.* 2025, Zhang *et al.* 2026). Hence, we argue that the geospatial transferability of human mobility generation models is determined not only by model architectures, but also by the intrinsic geographic differences between the source (training) region and the target (testing or transferred) region.

In this study, we focus on the generation of human travel origin-destination (OD) flow data, which represent aggregated individual movements between census geographic origin-destination pairs (Luca *et al.* 2021), with four representative mobility generation models, including *DeepGravity* (Simini *et al.* 2021), boosting regression tree (GBRT) (Robinson and Dilkina 2018), random forest (RF) (Pouerebrahim *et al.* 2019), and geo-contextual multitask embedding learning (GMEL) (Liu *et al.* 2020). The models are trained with geographic features and OD flows in a state. Subsequently, they are transferred to each county of the remaining states to evaluate their *geospatial transferability*. Meanwhile, we propose two domain shift metrics, *mutual information (MI) shift* and *Moran’s I spatial shift* (i.e., Moran shift for simplicity in the following text), to quantify the intrinsic differences in geographic feature distribution and spatial structure between a source area and a target area. Inspired by domain adaptation and generalization theories in transfer learning (Zhuang *et al.* 2020, Liu *et al.* 2023), we refer to these differences as *geographic domain shift*. Furthermore, we analyze the associations between geographic domain shift and model transferability to validate the proposed metrics and to understand how intrinsic differences in geographic features influence the transfer performance of human mobility OD flow generation models.

Notably, rather than developing a new geographically transferable deep learning model for generating human mobility OD flows, this study focuses on characterizing the intrinsic properties of geographic features in the training (source) and testing (target) regions *prior to model training*. By using a large-scale benchmark dataset of census tract-level commuting OD flows within each of 2,265 counties in the United States (Rong *et al.* 2025), we demonstrate the feasibility and effectiveness of the proposed geographic shift metrics. These metrics will enable the adaptive selection of suitable training areas for transferring a human mobility generation model to other geographic areas. Moreover, they are promising for guiding the optimization of

human mobility generation models toward improved geospatial transferability and fairness in synthetic human mobility data. The primary contributions of this study are three-fold:

- We conceptualize geographic domain shift and propose quantitative measures to characterize the intrinsic differences between source and target geographic domains. These measures provide a basis for interpreting geospatial transferability and may inform future model optimization and transfer learning strategies.
- We demonstrate that the geospatial transferability of multiple human mobility generation models varies substantially across target geographic areas. This finding highlights the need for more transparent reporting of transfer learning settings and for quantitative descriptions of how difficult a given geographic area can be transferred to or from.
- We employ linear mixed-effects regression to analyze the associations between the geographic domain shift metrics and model transferability. Our analysis results reveal substantial spatial heterogeneity and asymmetry in cross-region transferability.

2. Related work

2.1. *Geospatial transferability of human mobility generation models*

Human mobility generation can basically be grouped into individual mobility trajectory generation and aggregated mobility flow generation (Luca *et al.* 2021). In this study, we focus on the latter, which assigns inflow and outflow volumes to each pair of origin-destination (OD) geographic areas, referred to as OD flow generation. The geospatial transferability of a mobility generation model describes its ability to generate accurate mobility flows in a new or previously unseen geographic area. It is commonly based on OD flow distribution similarity measures, such as common part of commuters (CPC) and Jensen-Shannon Divergence (JSD), and magnitude-based errors like root mean squared errors (RMSE) of OD flows (Barbosa *et al.* 2018, Rong *et al.* 2025).

Existing studies on OD flow model transferability are largely *model-centric*, focusing on developing new mobility generation architectures followed by post-hoc evaluations of their transfer performance. For example, Simini *et al.* (2021) evaluated the transferability of the DeepGravity model across cities using a *leave-one-city-out* strategy. A growing body of work further explores spatial transfer learning for human mobility generation by learning domain-invariant representations that generalize across regions. These approaches include hierarchical mobility knowledge transfer (He *et al.* 2020), embedding the learning of novel data (Jiang *et al.* 2021), adversarial training (Rong *et al.* 2023a, Yuan *et al.* 2025), meta-learning (Wang *et al.* 2024b), and large language models (Yu *et al.* 2024). Although these studies validate their models using unseen geographic areas, they rarely investigate why the same model exhibits varying levels of geospatial transferability across different target regions. In particular, little

attention has been paid to how intrinsic geographic differences between regions influence the transfer performance of OD flow generation models.

2.2. Quantification of domain shift

As implied by prior work on transfer learning, a primary challenge hindering the geospatial transferability of deep learning models, which is independent of tasks, is *domain shift* across geographic areas (Pan and Yang 2009, Moreno-Torres *et al.* 2012, Liu *et al.* 2023). Domain shift typically encompasses both covariate shift and concept shift. The former refers to changes in the distribution of input features (Cai *et al.* 2025), while the latter denotes changes in the conditional distribution of labels given the input features (Garg *et al.* 2020) or changes in concept’s meanings over time (Shi *et al.* 2025). Since this study focuses on transferring a human mobility generation model to previously unseen geographic areas, where the label distribution is unknown, we mainly investigate the quantification of *covariate shift*.

In the field of transfer learning, many distance measures have been proposed to quantify the covariate shift (Farahani *et al.* 2021), such as mutual information (Blitzer *et al.* 2007), maximum mean discrepancy (MMD; Gretton *et al.* 2006), Wasserstein distance (Panaretos and Zemel 2019), correlation alignment (CORAL; Sun *et al.* 2017), and Kullback-Leibler (KL) divergence (Kullback and Leibler 1951). They primarily characterize discrepancies in the statistical distributions of covariates or feature representations.

From a geographic perspective, however, domain shift is often linked to *geographic similarity* (Goodchild 2004, McIntosh and Yuan 2005, Zhu *et al.* 2018). Methodologically, many geographic similarity metrics are derived from the statistical properties of covariate distributions (Zhu *et al.* 2015, Zhao *et al.* 2025a). Beyond purely statistical similarity, researchers have also modeled the *semantic similarity* of geographic features to distinguish between geographic domains (Schwering 2008, Janowicz *et al.* 2011). Nevertheless, these geographic similarity metrics generally do not account for the spatial distribution of geographic features and the associated structural differences in geographic phenomena across areas. Such spatial structure shifts are fundamental to geographic processes and can be explicitly characterized using classical spatial statistics, such as Moran’s I (Moran 1950, Lee and Li 2017).

2.3. Research gaps

Existing research on the geospatial transferability of human mobility generation models has largely focused on post-hoc evaluations to validate newly proposed models. In contrast, little attention has been paid to the intrinsic characteristics of cross-regional differences, such as geographic features and spatial structures, that underlie observed variations in transfer performance. Moreover, few studies have systematically examined how model transferability relates to the intrinsic characteristics of input data.

Meanwhile, although research on domain adaptation in machine learning has produced several methods for quantifying domain shifts, including geographic domains, they rarely account for changes in the spatial structures underlying geographic features or covariates. Therefore, a quantification framework for geographic domain shift, that is, capturing both distributional shifts and spatial-structure shifts, remains underexplored. This gap further constrains our understanding of the potential influence of intrinsic input data differences across locations and regions on the geospatial transferability of human mobility generation models.

3. Data and methods

3.1. OD mobility flow dataset

In this study, we employed an open large-scale OD mobility flow generation benchmark dataset¹, *CommutingODGen* (Rong *et al.* 2025), which provides both input geographic features and output OD flows, to investigate the geospatial transferability of a deep learning-based human mobility generation model. The rationale for using this dataset is twofold:

- (1) First, the OD flows span a wide range of geographic areas, covering 2,265 counties across 48 states in the United States², as shown in Figure 1 (a). This broad spatial coverage enables the systematic cross-region evaluation. Notably, the OD flows are restricted to within-county interactions and mainly depict commuting-related spatial interactions among census tracts.
- (2) Second, the dataset includes both input geographic features and corresponding OD flows for the year 2018. Specifically, the input geographic features have 97 demographic features and 36 categories of points of interest (POIs), which provide a comprehensive representation of the socioeconomic and built-environment characteristics underlying commuting mobility flows. As these geographic features are well aligned with each census tract within a county, additional alignment effort between input and output is unnecessary, thereby facilitating reproducible analysis.

3.2. Quantification of geographic domain shift

Unlike previous studies, which typically assumed that the training geographic area (i.e., the source geographic domain) is inherently different from the transferring geographic area (i.e., the target geographic domain) when assessing model transferability, we hypothesize that the actual difference between the source and target geographic domains varies across domain pairs. Accordingly, we propose to explicitly quantify geographic domain shift based on the intrinsic characteristics of geographic features and spatial structure used to train a human

¹<https://github.com/tsinghua-fib-lab/CommutingODGen-Dataset>

²Although the dataset includes 3233 counties from 50 states and Washington D.C. of the United States, we excluded the states of Alaska and Hawaii as well as Washington D.C., and kept 48 states in the Contiguous U.S. for our data analysis.

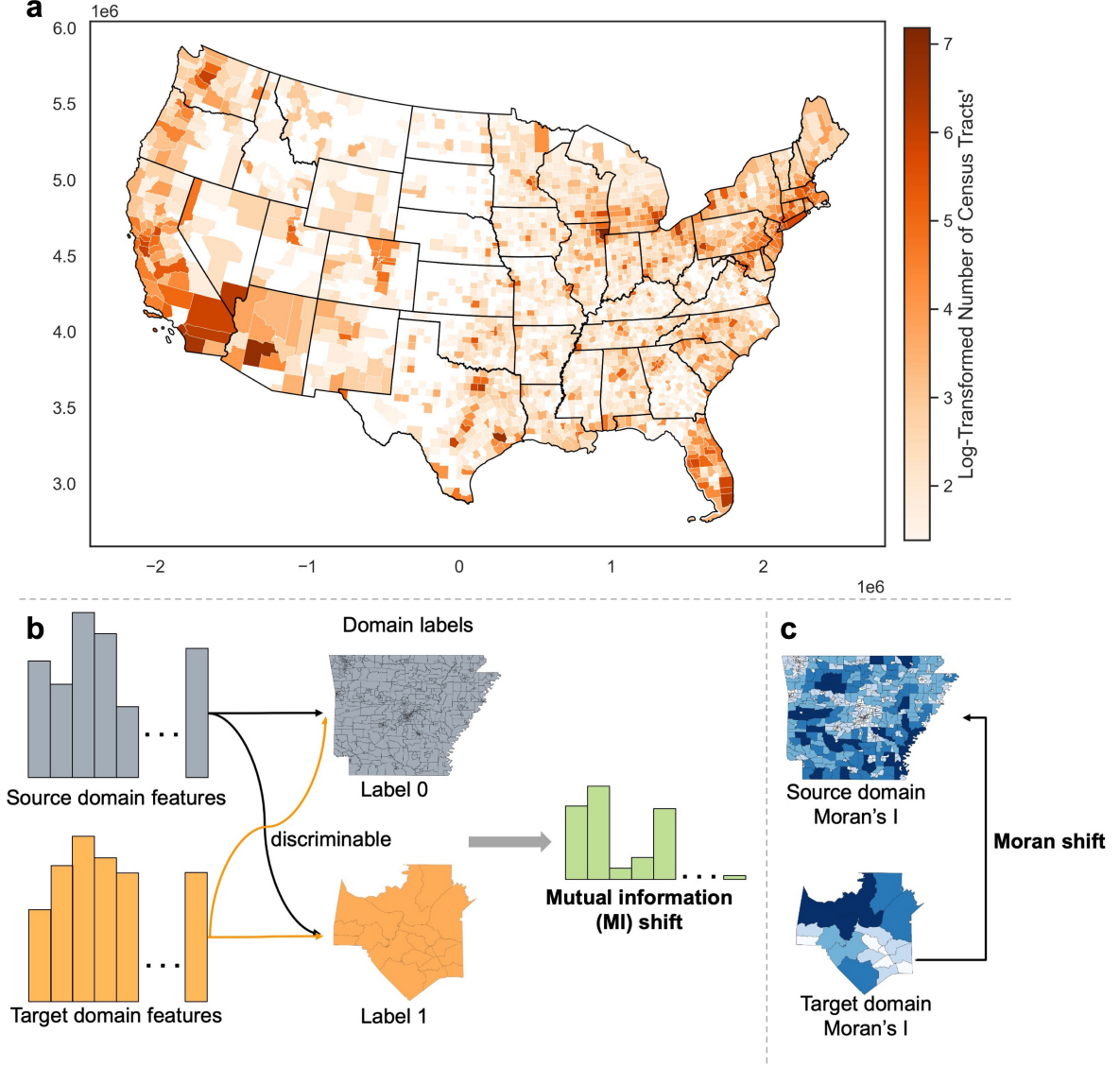


Figure 1. **a.** The county occupation of the used census tract-level commuting OD flow dataset in the United States. The color saturation denotes the log-transformed census tract number. **b.** The mutual information (MI) shift between input features of the source domain and input features of the target domain. **c.** The proposed Moran spatial shift between Moran’s I index of the target domain (I_{tar}) and that of the source domain I_{src} , i.e., $I_{tar} - I_{src}$.

mobility generation model. Specifically, we introduce *mutual information (MI) shift* to measure differences in the statistical distributions of geographic features between source and target domains, following classical formulations of covariate shift (Blitzer *et al.* 2007). In addition, to account for changes in spatial structure that may not be captured by purely distributional measures, we further propose a *Moran spatial shift*, derived from the spatial autocorrelation metric, Moran’s I (Moran 1950). Together, these two metrics characterize complementary aspects of geographic domain shift: MI shift captures feature-level (attributes) distributional discrepancies, while Moran shift reflects differences in the spatial dependence of features.

Notably, the quantification of geographic domain shift depends solely on the underlying

geographic features, such as the 97 demographic variables and 36 POI categories used in this study, and is independent of any specific human mobility generation model. This model-agnostic design enables the proposed metrics to be applied prior to model training and to support quantitative assessments of transferability across geographic domains.

Definition 3.1 (Mutual information shift). It measures the extent to which the N-dimensional geographic features can reduce uncertainty in distinguishing between the source and target geographic domains.

As illustrated in Figure 1 (b), given the i th dimension of N-dimensional geographic features X_i from both source distribution ($p_{src}(X_i)$) and target distribution ($p_{tar}(X_i)$) and domain labels,

$$Y = \begin{cases} 0, & \text{if } X_i \sim p_{src}(X_i) \\ 1, & \text{if } X_i \sim p_{tar}(X_i) \end{cases} \quad (1)$$

The mutual information shift score for the i th dimension of features is calculated as follows (Cover 1999),

$$MI_i(X_i; Y) = \sum_{x \in X_i} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2)$$

As a consequence, the overall MI shift based on N-dimensional geographic features is scored by averaging the sum of the N-dimensional MI shift scores,

$$MI(X; Y) = \frac{\sum_{i=0}^N MI_i}{N} \quad (3)$$

The MI shift score essentially measures the dependence of covariates on two geographic domains. A higher MI shift score means that the covariate is more domain-dependent (i.e., knowing the values of features X gives a lot of information about which geographic domain the data come from), indicating a larger shift between the source and target domains.

Definition 3.2 (Moran spatial shift). As shown in Figure Figure 1 (c), it measures the change in the spatial dependence of geographic features between the source and target domains (i.e., changes in spatial autocorrelation patterns), reflecting how features are more spatially clustered or dispersed from the source to the target domain.

The intuition for the Moran shift metrics is that existing geographic similarity metrics, which are mainly derived from classical distribution distances, such as cosine similarity and Euclidean distance of geographic features, rarely account for the spatial structures of geographic features (Zhao *et al.* 2025a), while most of the geographic phenomena, including human mobility flows (Boucherie *et al.* 2025, Zhao *et al.* 2025b), hold significant spatial dependence and spatial

heterogeneity (Goodchild and Li 2021). Given N-dimensional geographic features of the source and geographic domains, the corresponding multivariate Moran’s I indices (Lin 2023), I_{src} and I_{tar} , are calculated, respectively. Specifically, the standard formulation of a multivariate Moran’s I is represented as,

$$I = \frac{\sum_{i=1}^M \sum_{j=1}^M k^2 w_{ij} X_i^\top X_j}{\sum_{i=1}^M X_i^\top X_i} \quad (4)$$

Where M is the total number of geographic areas, $X_i = (x_{i1}, \dots, x_{in})$ represents N-dimensional geographic features at the i_{th} area, and w_{ij} is a spatial weight between i_{th} and j_{th} areas, which fulfills $\sum_{j=1}^M w_{ij} = 1$ for each geographic area.

Hence, the Moran spatial shift from the source to the target geographic domain is calculated as:

$$Moran_{src \rightarrow tar} = I_{tar} - I_{src} \quad (5)$$

Given this formula, the derived Moran shift score holds the sign: when it is > 0 , it indicates that the covariates in the target geographic domain are more spatially clustered than those in the source domain, and when it is < 0 , it indicates that the covariates in the target domain are more spatially dispersed. And Moran shift = 0 indicates the same overall strength of spatial autocorrelation (clustered, random, or dispersed) between source and target domains, but it does not necessarily show identical spatial patterns between source and target domains in terms of location of clusters and data values.

3.3. Training of human mobility generation model

In a data-driven human mobility generation, the choice of training and testing data is critical for revealing cross-region model performance, which in turn substantially influences the bias and fairness of the synthesized mobility data (Schlosser et al. 2021, Nijs et al. 2025, Liu et al. 2025). This motivates us to examine the relationships between intrinsic data characteristics and modeltransferability of human mobility generation models, beyond the extensively studied improvements achieved through model architecture design.

To this end, we chose a popular human mobility generation model, DeepGravity (Simini et al. 2021), as the generator of census tract-level commuting OD flows within each of 2265 counties. DeepGravity is inspired by the classical human mobility gravity mechanism and formulates the mobility flow generation process as a multinomial logistic regression problem, conditioned on the geographic features of an origin, its distances to potential destinations, and the known total outflow. Its core architecture consists of a feed-forward neural network with 15 hidden blocks, each of which is composed of a linear layer and a LeakyReLU activation layer³. Owing to its

³<https://github.com/scikit-mobility/DeepGravity>

strong performance and widespread adoption, DeepGravity has become a popular baseline for human mobility flow generation (Rong *et al.* 2025, Wang and Zhu 2025, Xu *et al.* 2025b).

Following the *leave-one-city-out* strategy performed by Simini *et al.* (2021), we adopted a *leave-one-state-out* experimental design to evaluate geospatial transferability of the DeepGravity model. Specifically, for each experiment, data from a single state were used for training and validation, while data from all remaining states were used for testing transfer performance. Within the selected training state, counties were used as the basic geographic units to preserve spatial coherence: 70% of counties were randomly assigned to the training set and the remaining 30% to the validation set for hyperparameter tuning. As a result, 48 separate models were trained, one for each state⁴. Given the large scale of the dataset, the batch size for training was set to 512 and the learning rate to 3×10^{-4} . Models were trained for up to 100 epochs, with early stopping applied when the validation loss increases for more than 10 consecutive epochs, in order to mitigate overfitting.

To verify the representativeness of DeepGravity in generating human mobility flows and examine the robustness of the associations between the proposed geographic domain shift metrics and the geospatial transferability of a human mobility generation model, we further trained two classical machine learning-based mobility generation models, boosting regression tree (GBRT) (Robinson and Dilkina 2018) and random forest (RF) (Pourebrahim *et al.* 2019), and one graph neural network-based mobility generation model, geo-contextual multitask embedding learning (GMEL) (Liu *et al.* 2020). While the parameters of these models follow the same settings as reported in (Rong *et al.* 2025), we solely changed the *max_depth* of GBRT from None to 3 to speed up the model training.

3.4. Evaluation of geospatial transferability

The evaluation of geospatial transferability is to compare the generated census tract-level OD flows T_{ij} with the ground truth commuting flows $\hat{T}_{ij}$ within each target domain testing county. Two metrics were used to evaluate the geospatial transferability of the model trained in a source state domain: the common part of commuting (CPC) and the root mean squared error (RMSE), as calculated by Equation (6) and Equation (7).

$$CPC = \frac{2 \sum_{i,j} \min(T_{ij}, \hat{T}_{ij})}{\sum_{i,j} T_{ij} + \sum_{i,j} \hat{T}_{ij}} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{ij} (T_{ij} - \hat{T}_{ij})^2} \quad (7)$$

In the context of OD flow generation, the CPC metric measures the overlap between the

⁴Washington, D.C. was excluded due to insufficient data to support a 7:3 train-validation split.

observed and generated flows, that is, the smaller of the two values. Then it sums the overlap across all OD pairs and normalizes it by total flow volume. It captures the overall spatial structure similarity between the observed and generated flows. A high CPC means the model captures where commuters actually go from origins, while a low CPC means that the model misses important flow patterns and allocates flows to incorrect destinations. In terms of RMSE, it measures the average squared difference between the observed and generated flows, that is, the average magnitude of prediction error. It mainly reflects numerical accuracy more than the structural overlap of mobility flows that are depicted by the CPC metric.

3.5. *Linear mixed-effects regression model*

Furthermore, we want to explore the associations between proposed geographic domain shift metrics and the geospatial transferability of the OD flow mobility generation model. In our study, the geospatial transferability of DeepGravity and the other three models, measured by CPC and RMSE, is typically regarded as a response to the geographic domain shift (measured by MI shift and Moran shift). Because each county was evaluated repeatedly under different source training states, the resulting CPC and RMSE observations were not independent. To account for this repeated-measure structure and unobserved county-level heterogeneity, we included a county-specific random effect. Source training states were modeled as fixed effects because they represent a finite set of experimental conditions under which geospatial transferability was evaluated. Accordingly, a linear mixed-effects regression model was chosen to model the relationship between model transferability and geographic domain shift scores, including MI shift and Moran shift, which is formulated as:

$$y_{c,s} = \beta_0 + \beta_1 MI_{c,s} + \beta_2 Moran_{c,s} + \gamma_s + \delta_c + \epsilon_{c,s} \quad (8)$$

Where c is target county domain and s is a source training state domain, $y_{c,s}$ denotes a transferability metric (CPC or RMSE) in c under the condition of s , β_0 is a global intercept, and β_1 and β_2 represent the population-average MI shift effect and Moran shift effect. Moreover, γ_s is the fixed effect of training state (dummy variable), δ_c is a county-specific random intercept (which draws from $N(0, \sigma_s^2)$, accounting for unobserved county-level heterogeneity or variation), and $\epsilon_{c,s}$ denotes the residual error. Notably, since the statistical distribution of RMSE is right-skewed and the value range of CPC is $[0,1]$, we performed a log transformation of the RMSE and a logit transformation of CPC for the regression analysis. The implementation of the linear mixed-effects model (MixedLM) was using the *statsmodels* library⁵ in Python.

⁵https://www.statsmodels.org/stable/examples/notebooks/generated/mixed_lm_example.html

4. Results

4.1. Predictive performance of human mobility generation models

The predictive performance of trained mobility generation models is evaluated on the unseen regions from the source training state. As reported in Table 1, DeepGravity outperforms the other three human mobility generation models, achieving the highest CPC and the lowest RMSE. Notably, the high standard deviations of RMSE values from all four models indicate that there are extreme cases in the predictions of numerical values of OD flows. Given such a comparative performance, we chose the DeepGravity model as the primary focus in our research.

Table 1. The predictive performance of different mobility generation models

Models	CPC		RMSE	
	Mean	Std Dev	Mean	Std Dev
DeepGravity	0.603	0.101	82.533	61.706
RF	0.558	0.115	90.061	68.789
GBRT	0.525	0.117	93.515	70.256
GMEL	0.464	0.164	146.94	97.759

4.2. Characteristics of geospatial transferability of mobility generation

Spatial heterogeneity of transferability. Figure 2 (a) and (b) illustrate the transfer performance of the DeepGravity model trained with different source states, measured by CPC and RMSE, respectively. The achieved overall CPC and RMSE by the DeepGravity model hold significant spatial heterogeneity, as shown in the varying saturation of colors in the figures. It is worth noting that CPC generally captures global mobility flow patterns, whereas RMSE emphasizes absolute flow magnitudes at individual spatial units and is more sensitive to outliers. Therefore, the two metrics show distinctive spatial patterns.

Specifically, DeepGravity achieved a high CPC (over 0.70) mainly in the Midwestern states of U.S., i.e., dark blue areas in Figure 2 (a). In contrast, in those areas along the east and east coasts, the mean CPC values of DeepGravity models trained with all of the individual source states are much lower, e.g., in the range of [0.33, 0.48] for light blue areas, which indicates a more complex human mobility flow structure that cannot be easily learned from other states. Meanwhile, those tall hexagon cubes indicate the larger standard deviation of CPC by different individual source state-trained models, indicating inconsistent performance of the same deep learning architecture (i.e., DeepGravity in this study) across source training states. Geographically, these areas of high CPC standard deviation are also areas of high mean values of CPC.

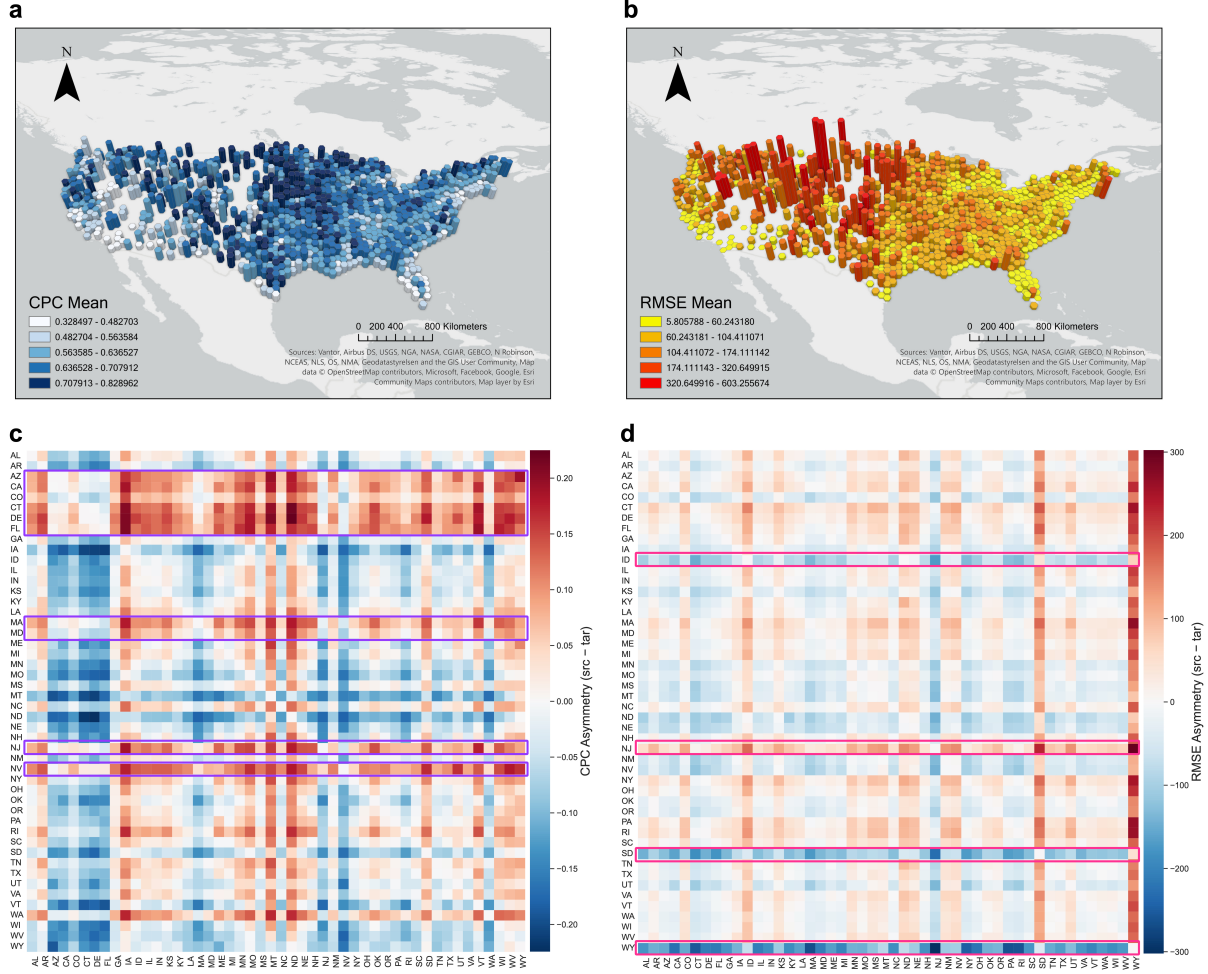


Figure 2. The hexagon cube maps of OD flow generation transfer performance, **a.** CPC and **b.** RMSE, which were aggregated with hexagon cells across the 2265 counties from 48 states, where the counties of a state construct the source domain for training, and the remaining counties are transferred to. The color saturation of a hexagon cube represents the mean values of CPC or RMSE achieved in target counties within the hexagon area, while the height of the hexagon cube denotes the standard deviation of CPC or RMSE achieved by DeepGravity models trained with different source states in the same hexagon area. Notably, the height in **a.** is scaled up to 5000 times the CPC std., and the height in **b.** is scaled up to 10 times the RMSE std. The bidirectional evaluation of state pairs, one of which acts as a source state domain for training and the other one is a target domain for transfer, shows the asymmetry of geospatial transferability regarding **c.** CPC and **d.** RMSE. Each row represents the geospatial transferability from a source state to the other states.

Regarding the RMSE-based transfer performance, the generated OD flows by DeepGravity models in the Western and Midwestern U.S. states held substantial errors, and the change of the individual source states for training resulted in considerable changes in RMSE, as indicated by the high saturation of orange and tall heights of hexagon cubes. Furthermore, it is visually clear that the overall RMSE gradually decreases from the Midwestern to the Eastern states, indicating increased transfer performance, except for several areas in Maine and Florida. The distinct patterns captured by CPC and RMSE metrics reflect the different nature of the measurement (i.e., global mobility structure v.s. absolute flow magnitude).

Asymmetry in geospatial transferability. For each DeepGravity model trained on a single source state, we aggregated CPC and RMSE values across target testing counties according to their corresponding states, thereby deriving state-to-state transfer performance. To examine whether a pair of states can transfer symmetrically between their respective domains, we constructed an asymmetry metric defined as the difference between transfer performance from the source state to the target state and that from the target state to the source state. Accordingly, the difference is directional and signed, where values greater than zero indicate that the DeepGravity model trained on the source state exhibits higher geospatial transferability to the target state than the model trained on the target state when evaluated on the source state.

Figure 2 (c) and (d) present the CPC and RMSE asymmetry matrices, respectively. Specifically, the CPC asymmetry of DeepGravity ranges from 0 to 0.2, with models trained on Arizona (AZ), California (CA), Colorado (CO), Connecticut (CT), Delaware (DE), Florida (FL), New Jersey (NJ), and Nevada (NV) generally achieving lower CPC when transferred to other states than models trained elsewhere, indicating their heterogeneous OD flow data distributions. In terms of RMSE, models trained on data from Idaho (ID), South Dakota (SD), and Wyoming (WY) tended to underestimate absolute flows in other states, whereas models trained on New Jersey (NJ) typically overestimated absolute origin–destination flows when applied outside the source state.

4.3. *Distribution of geographic domain shift*

The geographic domain shift from a source state domain (for training) to the target county domain (for evaluation) is quantified based on MI shift and Moran shift as introduced in Section 3.2. Notably, since the used benchmark dataset only provides a county-level multivariate distance matrix and misses the raw geographic information of census tracts for OD flows, we can only calculate the county-level Moran’s I index, and then compute the average Moran’s I of multiple counties in a state to denote the source state’s Moran’s I. After that, the source-target Moran spatial shifts are calculated according to Equation (5).

Figure 3 (a) visualizes their joint distribution of MI shift and Moran shift. While the MI shift of domain pairs ranges from 0.00 to 0.14 generally, the majority stay in the range of [0.00, 0.06]. Regarding Moran shift, it changes between -0.75 and 0.6 approximately, mostly staying in the range of [-0.4, 0.2]. Notably, a positive sign of Moran shift indicates that the spatial structures of geographic features in the target county domain are typically more clustered than the spatial structures shown in the source state domain, while a negative sign indicates more dispersed spatial structures. Moreover, the shape of the joint distribution of MI shift and Moran shift reveals that Moran shift varies even though MI shift is very minor, e.g., around 0.00, implying the non-collinear relationship of the proposed Moran shift with MI shift. The two metrics capture

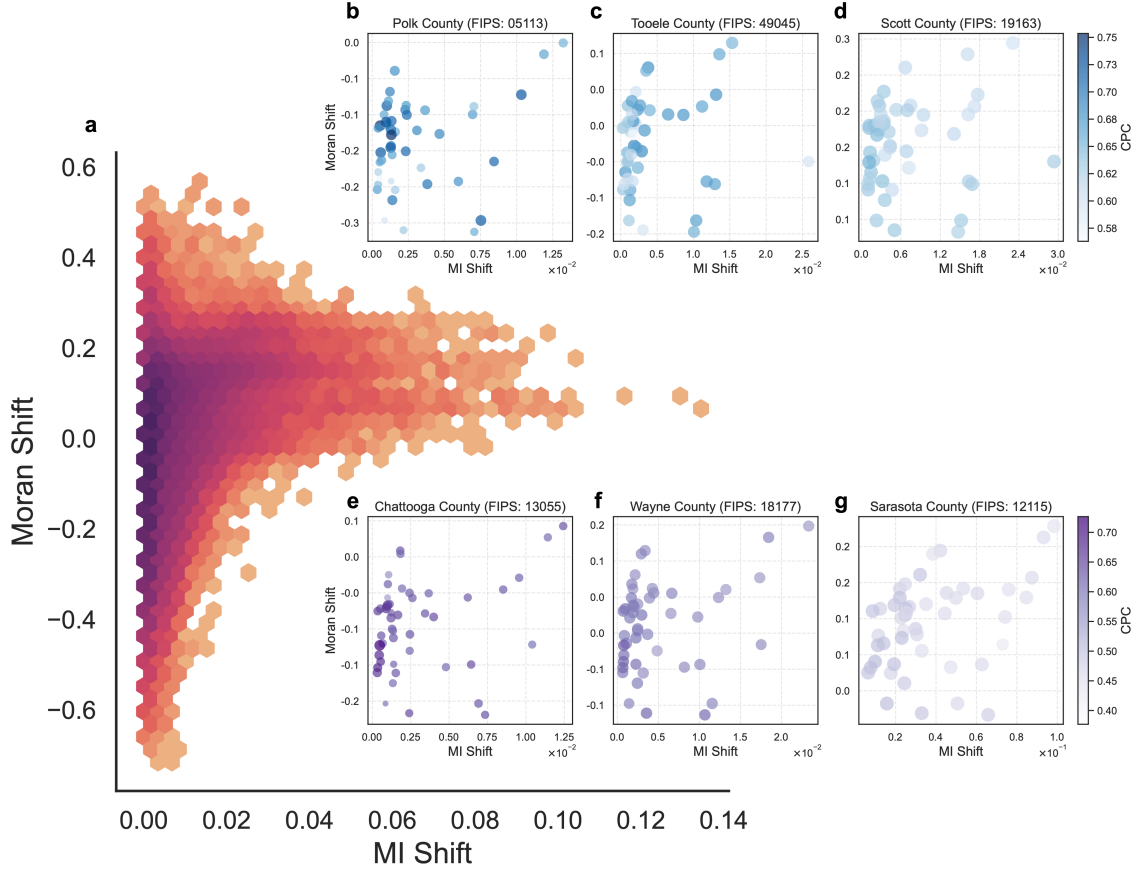


Figure 3. **a.** The joint distribution of mutual information (MI) shift (x-axis) and Moran’s I spatial shift (y-axis). **b-d.** The MI-Moran’s I shift distribution of different reference states in three counties, where the MI shifts remain at the 25%, 50%, and 75% quantiles in ascending order, respectively, and **e-g.** The distribution in three counties whose Moran’s I index shifts remain at the 25%, 50%, and 75% quantiles, respectively. The saturation of color denotes the CPC transferability, with darker indicating a higher CPC. The size of dots represents the RMSE transferability, with the larger size meaning a larger RMSE.

different aspects of the geographic domain shift.

Additionally, we depicted the joint distributions of three counties based on the quantiles of MI shifts (Figure 3 b-d) and the quantiles of Moran shifts (Figure 3 e-g), respectively, to illustrate the details of the geospatial transferability evaluations at a county. In each scatter figure, 47 out of 48 source states (excluding the state containing the transferred state) result in 47 Moran shift and MI shift from their source domain to the same target county domain, and the corresponding transferability of the trained DeepGravity models. Generally, the dots in each figure are dispersed along both MI shift and Moran shift axes, indicating that the geographic domain varies with the source state domains, given the same target county domain. Among the three counties with an increase of MI shift (Figure 3 b-d), from Polk County in Arkansas, Tooele County in Utah, to Scott County in Iowa, their Moran shift generally increases, and its scatter ranges are from $[-0.3, 0.0]$ to $[-0.2, 0.1]$ and $[0.1, 0.3]$.

In terms of the Moran shifts, the 25% quantile is Chattooga County in Georgia, the 50%

quantile is Wayne County in Indiana, and the 75% quantile is Sarasota County in Florida. In particular, the corresponding MI shift of Wayne County (Figure 3f, at the 50% quantile Moran shift level) span over the largest range, [0.0, 2.5]. Additionally, the color saturation and size of dots across target county domains demonstrate that the DeepGravity models typically perform better under lower MI shift (e.g., Figure 3b) and Moran shift (e.g., Figure 3d) conditions. However, within each target county, the variance of transferability caused by the change of source state domains cannot be overlooked, which is a detailed illustration of high cubes in 3D hexagon choropleth maps as shown in Figure 2.

4.4. Associations of geographic domain shifts with geospatial transferability

To understand the associations of geographic domain shifts with the geospatial transferability of individual source state-trained models at different target county domains, we fitted the linear mixed-effects regression models (see Methods Section 3.5) with CPC and RMSE on target counties for the four trained models (i.e., DeepGravity, RF, GBRT, and GMEL), respectively. As reported in Table 2, MI shift is negatively associated with CPC, indicating improved transferability under smaller mutual-information shift, whereas Moran shift exhibits a positive association with CPC, suggesting improved geospatial transferability when the spatial autocorrelation pattern in the target domain is stronger than that in the source domain. According to the regression for log-transformed RMSE, we found that MI shift is positively associated with $\log(\text{RMSE})$, indicating increased RMSE, i.e., poorer transfer performance, under larger mutual-information shift. In contrast, Moran shift exhibits a negative association with $\log(\text{RMSE})$, suggesting smaller RMSE, i.e., improved geospatial transferability, with stronger spatial autocorrelation pattern in the target domain.

Table 2. Linear mixed-effects regression results for the *CPC (Logit-transformed)* response.

Models	Predictors	CPC (Logit-transformed) response		RMSE (Log-transformed) response	
		Coefficients	p-value	Coefficients	p-value
DeepGravity	MI shift	-0.023	<0.001	0.010	<0.001
	Moran shift	0.013	<0.001	-0.007	<0.001
RF	MI shift	-0.357	<0.001	0.239	<0.001
	Moran shift	0.178	<0.001	-0.097	<0.001
GBRT	MI shift	-0.430	<0.001	0.287	<0.001
	Moran shift	0.175	<0.001	-0.098	<0.001
GMEL	MI shift	-0.213	<0.001	0.149	<0.001
	Moran shift	0.071	<0.001	-0.058	<0.001

Since a higher CPC refers to better geospatial transferability but RMSE is the opposite, the regression analysis results for both transferability metrics jointly imply that a higher MI shift (i.g., more distinctive feature distributions) is associated with a poorer transfer performance of a mobility generation model, while a stronger spatial autocorrelation can inform a better transferability. Moreover, the signs of associations between the proposed two geographic domain shift metrics and the two transferability metrics are consistent across all the four trained mobility generation models, which suggests the generalizable associations independent of specific human mobility generation models.

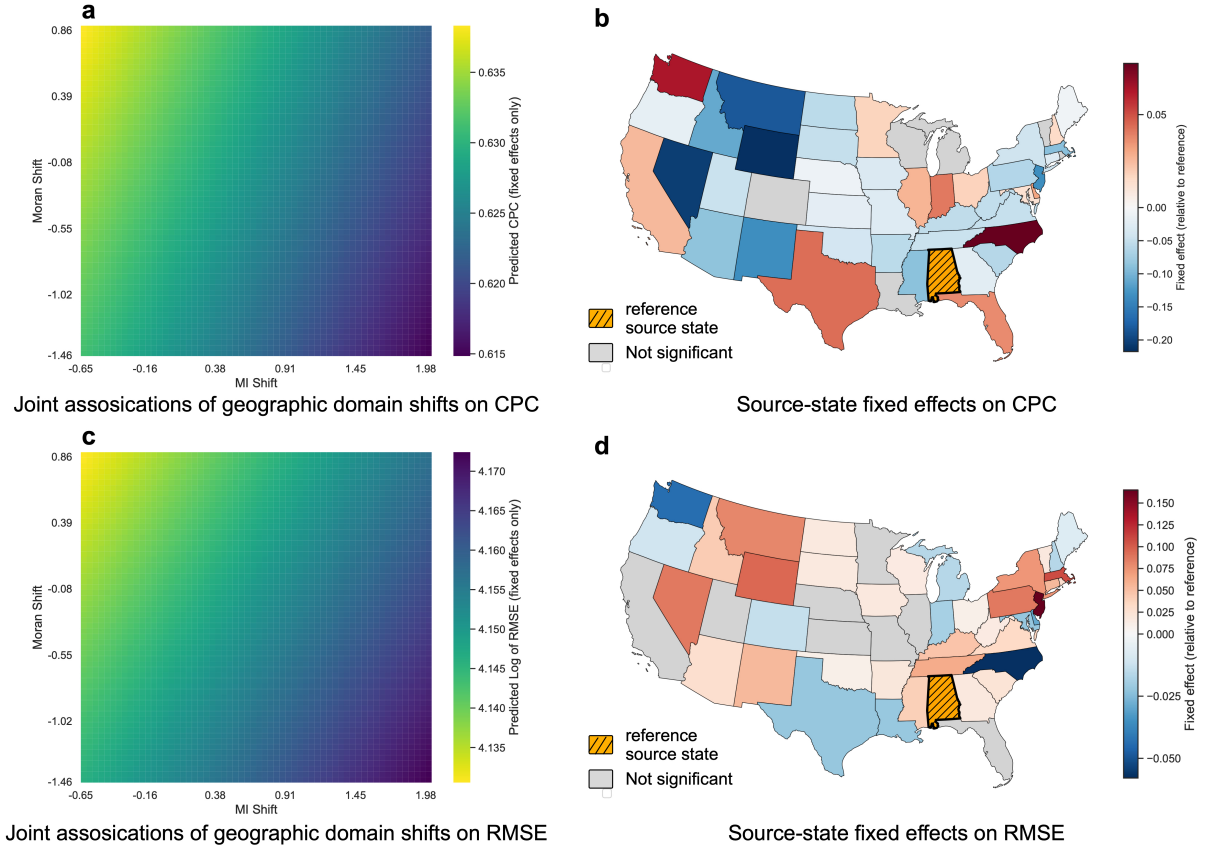


Figure 4. The joint associations of MI shift and Moran shift on CPC in (a) and on RMSE in (c) of fitted linear mixed-effects models based on DeepGravity. The derived fixed effects of source (training) states are mapped in (b) and (d), respectively.

Additionally, the joint associations of MI shift and Moran shift with geospatial transferability of DeepGravity in fitted models are visualized in Figure 4 (a) and (c) regarding CPC and RMSE, respectively. Moreover, the source-state fixed effects were mapped in Figure 4 (b) and (d), where 6 out of 48 source states did not show statistically significant fixed effects on CPC, while 8 source states were not statistically significant in the regression of log-transformed RMSE.

5. Discussion

5.1. *Biased evaluations of geospatial transferability in human mobility generations*

The systematic leave-one-state-out evaluations in Section 4.2 demonstrate that geospatial transferability exhibits substantial spatial heterogeneity and asymmetry across state pairs, regardless of whether CPC or RMSE is used. These findings indicate that the transfer performance of a human mobility generation model is biased by the choice of source-domain training data and the target-domain transfer setting. However, existing human mobility generation studies (Rong *et al.* 2023a, 2025, Li *et al.* 2025) typically rely on a small number of regions and averaged overall evaluation metrics (e.g., mean CPC and mean RMSE) to compare models and assess cross-region generalization. While this evaluation practice is reasonable and facilitates straightforward performance comparisons across different deep learning-based mobility generation models, it risks obscuring the models’ transferability across geographic settings..

Specifically, changes in the source-domain training data or the target-domain testing region may lead to substantial performance variation, making it uncertain whether an initially well-performing model will maintain its effectiveness across regions. As illustrated in Figure 2, the mean CPC of DeepGravity models trained on different individual source states ranges approximately from 0.58 to 0.65, yielding a difference of about 0.07. Notably, some existing studies report CPC differences between baseline models that are smaller than this range, which might confound whether the observed performance differences arise from model architectures or from the intrinsic characteristics of the datasets.

Additionally, we observe that transfer performance assessed using CPC is not always consistent with that evaluated using RMSE. In other words, these regions with higher CPC are prone to getting larger RMSE. However, this discrepancy is not contradictory, because CPC captures global mobility flow patterns and their overall spatial structure similarity, while RMSE emphasizes absolute flow magnitudes at individual spatial units and is therefore more sensitive to outliers (Rong *et al.* 2024). Such an inconsistency indicates that geospatial transferability is a concept of multiple dimensions and may be constrained by geographic scales. As a consequence, different evaluation metrics can reflect different aspects of geospatial transferability. Furthermore, it highlights a fundamental challenge in human mobility generation, that is, how to make a trade-off between accurately reproducing global patterns and capturing fine-grained local details (Mauro *et al.* 2022).

5.2. *Asymmetric geospatial transferability*

As illustrated in Figure 2 (c) and (d), the geospatial transferability, as evaluated by both CPC and RMSE, exhibits substantial asymmetry. Specifically, if an OD flow generation model trained in a source region can generate realistic OD flows in a target region, the converse does not necessarily hold: a model trained in the target region typically does not achieve equivalent transferability to the source region. Although such asymmetry in transferability has been observed in many mobility generation studies (He *et al.* 2020, Rong *et al.* 2023a) and other geospatial data generation studies (Yu *et al.* 2025), this study is the first to formally conceptualize it. This asymmetry is likely driven by heterogeneous geographic structures between source and target regions, whereby models trained in the source region may fail to capture the more complex relationships present in the target region.

More importantly, this transfer asymmetry further supports our rationale for introducing a signed metric, such as the Moran shift proposed in this study. Unlike conventional domain shift measures that capture solely distributional differences or similarities, such as mutual information shift, the Moran shift can explicitly represent the direction of transfer from the source region to the target region. Additionally, although classical theories of geographic similarity (Zhu and Turner 2022) are related to geographic domain shift, most similarity metrics are also directional (Yu *et al.* 2025). This further highlights the significance and timeliness of our proposed Moran shift as a new measure for quantifying geographic domain differences.

5.3. *Feasibility of the proposed geographic domain shift metrics*

Given the varying transfer performance across source-target domain pairs, we hypothesize that they are primarily influenced by the intrinsic data characteristics of datasets encompassing both geographic domains. Accordingly, this study proposes two metrics, mutual information (MI) shift and Moran shift, to capture intrinsic covariate shifts from the source to the target geographic domain. The MI shift is a classical metric for quantifying covariate shift in terms of distributional similarity, whereas the Moran shift is newly introduced to characterize directional changes in the spatial structure of geographic features between the source and target domains.

As demonstrated in Section 4.3, the newly introduced Moran spatial shift captures additional covariate shift information that is not characterized by the classical MI shift. Together with the MI shift, it is further shown to be indicative of model transferability, as indicated by the CPC and RMSE of human mobility generation, based on analyses using linear mixed-effects regression models in Section 3.5. These results, in turn, support our hypothesis that geographic domain shifts contribute to biased evaluations of geospatial transferability. Moreover, the MI shift and Moran shift may complement geographic similarity in explaining the cross-region model transferability for other geospatial tasks (Yu *et al.* 2025).

Additionally, geographic domain shift has received considerable attention in the earth observation and remote sensing communities (Kalluri *et al.* 2023, Al-Emadi *et al.* 2025a,b, Crasto 2025, Doerksen and Kerner 2026). As their primary geospatial tasks typically involve semantic and instance segmentation of satellite imagery, existing studies mainly focus on shifts in the distribution of geographic feature classes rather than their spatial structures. Therefore, the proposed geographic domain shift metrics remain valuable for providing deeper insights into the intrinsic characteristics of satellite image datasets.

5.4. Association between training dataset size and transferability

Since our research adopted the *leave-one-state-out* strategy to repeatedly evaluate the model transferability of mobility generation models on held-out target counties, the training dataset size, i.e., the number of census tracts acting as origins and destinations in our used mobility OD flow benchmark dataset, is inevitably not the same across states. Hence, it may act as a confounding factor in the association analysis of transfer performance with the introduced MI shift and Moran shift, given the common belief in scaling laws in deep learning, where the size of a training dataset fed into neural networks is one of the important drivers (Hestness *et al.* 2017, Cherti *et al.* 2023, Merchant *et al.* 2023).

Table 3. Ordinary least squares regression results of geospatial transferability with dataset size, mutual-information shift, and Moran shift. Since the mean CPC distribution is bounded to 0 and 1, it was logit-transformed in a linear regression.

Responses	Predictors	Coefficients	p-value
Mean CPC (Logit-transformed)	Log-transformed data records	0.015	0.171
	mean MI shift	-0.026	0.018
	mean Moran shift	0.003	0.674
Mean RMSE	Log-transformed data records	0.006	0.496
	mean MI shift	0.012	0.152
	mean Moran shift	-0.015	0.018

To examine the association between dataset size and transferability, we averaged the MI shift and Moran shift across held-out target counties, along with the corresponding CPC and RMSE. As a result, the mean MI shift value, the mean Moran shift value, and the census tract number act as the three intrinsic characteristics of a source-state training dataset. We conducted a regression analysis with the ordinary least squares method. As reported in Table 3, only the mean MI shift remains a significant negative association with the mean CPC (logit-transformed), while in the regression of mean RMSE, only the mean Moran shift holds a significant negative association. This indicates that the proposed geographic domain shift metrics may serve an essential role in influencing the overall transferability of a machine learning model.

5.5. *Implications of geographic domain shift for human mobility generation*

The broader implications of geographic domain shift for data-driven human mobility generation are twofold:

- (1) *Dataset curation.* Section 5.4 reveals a counterintuitive result: the training dataset size is not the primary driver of the transferability of our human mobility generation model, DeepGravity, compared with the MI shift and Moran shift. This finding suggests that dataset curation for human mobility generation should move beyond simply increasing data volume and instead prioritize intrinsic dataset characteristics, such as the proposed geographic domain shift metrics, to achieve higher geospatial transferability (Janowicz *et al.* 2025, Stewart *et al.* 2025). The finding links to the *dataset diversity* research in the general machine learning theory (Mandal *et al.* 2021, Aroyo *et al.* 2023, Zhao *et al.* 2024, Nougnanke *et al.* 2025). While many studies have mentioned geographic bias as one aspect of dataset diversity, the proposed metrics may be able to provide more insights about the dataset diversity from the perspective of spatial distribution.
- (2) *Human mobility generation model development.* The linear mixed-effects regression analysis reveals fixed effects of the source training state on transfer performance of the DeepGravity model, consistent with previously reported fixed effects of geographic areas (Chen *et al.* 2025). This finding motivates the explicit incorporation of spatial structure in developing geographically transferable deep learning models for human mobility generation. For example, aggregation of geographic information in convolutional architectures has been refined by accounting for spatial heterogeneity and autocorrelation of geographic phenomena (Deng *et al.* 2025, Guo *et al.* 2025). Furthermore, as improving transferability commonly involves reducing domain shifts through feature selection and representation learning (Pan and Yang 2009, Farahani *et al.* 2021), the newly introduced Moran shift metric, together with the MI shift, can be naturally integrated into model optimization to enhance the geospatial transferability of human mobility generation models and other task-specific deep learning models.

5.6. *Limitations and future work*

Despite the systematic investigation of the associations between intrinsic dataset characteristics and the geospatial transferability of human mobility generation models, this study has several limitations that warrant future work. First, our analysis relies on a single mobility flow dataset in one country, which may introduce partial dependence between the derived geographic domain shift metrics and the transferability measures (i.e., CPC and RMSE). Although we employ a linear mixed-effects regression model to account for fixed effects of source training states and random effects of target counties, more diverse mobility datasets (e.g., different geographic

scales, mobility types, and countries) and more themes of geospatial data generation tasks (e.g., population synthesis) are needed in future work (Moska *et al.* 2025) to derive more independent estimates of geographic domain shifts and transferability and to more directly examine their relationships. Second, the quantitative associations between geographic domain shift and geospatial transferability should be further validated using other human mobility generation models or algorithms, to assess whether the indicative power of the proposed MI shift and Moran shift generalizes across different modeling approaches. Moreover, additional metrics should be developed in the future to quantify a broader range of spatial structure changes (e.g., polycentric versus monocentric patterns) beyond the global Moran shift introduced in this study, which solely focuses on the overall spatial autocorrelation of geographic features.

6. Conclusion

This study systematically investigates the geospatial transferability of representative human mobility generation models using the large-scale *CommutingODGen* dataset, which covers census-tract-level origin–destination flows across 2,265 counties in 48 U.S. states. Motivated by the role of intrinsic dataset characteristics in shaping transfer performance across source and target domains, we propose two geographic domain shift metrics, mutual information shift and Moran shift, to quantify domain shifts within a dataset. We further employ a linear mixed-effects regression model to disentangle the associations between intrinsic dataset characteristics and the model transferability. The results reveal heterogeneous and asymmetric patterns of transfer performance between the source and target domains, and demonstrate the complementary roles of mutual information shift and Moran spatial shift in characterizing geographic domain shift. Furthermore, the regression analysis shows significant associations between the proposed domain shift metrics and the model transferability, demonstrating the feasibility of these metrics for assessing the intrinsic characteristic differences between training and testing geographic datasets. Therefore, these metrics have considerable potential to guide the selection of geographic training datasets for developing more geographically transferable human mobility generation models and other GeoAI models.

Data and code availability statement

The information about the data and code of this study is publicly available at GitHub: <https://github.com/GeoDS/GeoDomainShift-Mobility>.

Acknowledgements

Dr. Zhiyong Zhou sincerely acknowledges the Postdoc.Mobility Fellowship (No. 235381) funded by the Swiss National Science Foundation in supporting this research.

References

- Al-Emadi, S.A., Yang, Y., and Offi, F., 2025a. Analysing satellite imagery classification under spatial domain shift across geographic regions. *International Journal of Computer Vision*, 133 (11), 7672–7709.
- Al-Emadi, S.A., Yang, Y., and Offi, F., 2025b. Benchmarking object detectors under real-world distribution shifts in satellite imagery. *In: Proceedings of the Computer Vision and Pattern Recognition Conference*. 8299–8309.
- Aroyo, L., *et al.*, 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36, 53330–53342.
- Barbosa, H., *et al.*, 2018. Human mobility: Models and applications. *Physics Reports*, 734, 1–74.
- Blitzer, J., Dredze, M., and Pereira, F., 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *In: Proceedings of the 45th annual meeting of the association of computational linguistics*. 440–447.
- Boucherie, L., Maier, B.F., and Lehmann, S., 2025. Decoupling geographical constraints from human mobility. *Nature Human Behaviour*, 1–12.
- Cai, T., Namkoong, H., and Yadlowsky, S., 2025. Diagnosing model performance under distribution shift. *Operations Research*.
- Chen, Y., Ma, Q., and Tao, R., 2025. Addressing the fixed effects in gravity model based on higher-order origin-destination pairs. *International Journal of Geographical Information Science*, 39 (5), 1162–1182.
- Cherti, M., *et al.*, 2023. Reproducible scaling laws for contrastive language-image learning. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2818–2829.
- Cover, T.M., 1999. *Elements of information theory*. John Wiley & Sons.
- Crasto, R., 2025. Robustness to geographic distribution shift using location encoders. *arXiv preprint arXiv:2503.02036*.
- Deng, R., Li, Z., and Wang, M., 2025. Geoaggregator: An efficient transformer model for geo-spatial tabular data. *In: Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, 11572–11580.
- Doerksen, K. and Kerner, H., 2026. Earthshift: a benchmark for measuring robustness to real-world distribution shifts in earth observation. *arXiv preprint arXiv:2605.29330*.
- Farahani, A., *et al.*, 2021. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, 877–894.
- Gao, S., Hu, Y., and Li, W., 2023. *Handbook of geospatial artificial intelligence*. CRC Press.
- Garg, S., *et al.*, 2020. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33, 3290–3300.
- Gonzalez, M.C., Hidalgo, C.A., and Barabasi, A.L., 2008. Understanding individual human mobility patterns. *nature*, 453 (7196), 779–782.
- Goodchild, M.F., 2004. The validity and usefulness of laws in geographic information science and geography. *Annals of the Association of American Geographers*, 94 (2), 300–303.
- Goodchild, M.F. and Li, W., 2021. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118 (35), e2015759118.

- Gretton, A., *et al.*, 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.
- Guo, H., *et al.*, 2025. Regioncn: Spatial-heterogeneity-aware graph convolutional networks. *Annals of the American Association of Geographers*, 1–17.
- He, T., *et al.*, 2020. What is the human mobility in a new city: Transfer mobility knowledge across cities. *In: Proceedings of The Web Conference 2020*. 1355–1365.
- Hestness, J., *et al.*, 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.
- Hou, C., *et al.*, 2025. Transferred bias uncovers the balance between the development of physical and socioeconomic environments of cities. *Annals of the American Association of Geographers*, 115 (1), 148–166.
- Hou, X., *et al.*, 2021. Intracounty modeling of covid-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. *Proceedings of the National Academy of Sciences*, 118 (24), e2020524118.
- Huang, X., *et al.*, 2022. Staying at home is a privilege: Evidence from fine-grained mobile phone location data in the united states during the covid-19 pandemic. *Annals of the American Association of Geographers*, 112 (1), 286–305.
- Janowicz, K., *et al.*, 2025. Geofm: how will geo-foundation models reshape spatial data science and geoai? *International Journal of Geographical Information Science*, 39 (9), 1849–1865.
- Janowicz, K., Raubal, M., and Kuhn, W., 2011. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2, 29–57.
- Jiang, R., *et al.*, 2021. Transfer urban human mobility via poi embedding over multiple cities. *ACM Transactions on Data Science*, 2 (1), 1–26.
- Kalluri, T., Xu, W., and Chandraker, M., 2023. Geonet: Benchmarking unsupervised adaptation across geographies. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15368–15379.
- Kullback, S. and Leibler, R.A., 1951. On information and sufficiency. *The annals of mathematical statistics*, 22 (1), 79–86.
- Lee, J. and Li, S., 2017. Extending moran’s index for measuring spatiotemporal clustering of geographic events. *Geographical Analysis*, 49 (1), 36–57.
- Li, Y., *et al.*, 2025. Cross city traffic flow generation via retrieval augmented diffusion model. *In: The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Lin, J., 2023. Comparison of Moran’s I and Geary’s C in multivariate spatial pattern analysis. *Geographical Analysis*, 55 (4), 685–702.
- Liu, J., *et al.*, 2023. Towards out-of-distribution generalization: A survey. Available from: <https://arxiv.org/abs/2108.13624>.
- Liu, Z., *et al.*, 2025. Generating equitable urban human flows with a fairness-aware deep learning model. *Cities*, 167, 106296.
- Liu, Z., *et al.*, 2020. Learning geo-contextual embeddings for commuting flow prediction. *In: Proceedings*

- of the AAAI conference on artificial intelligence. vol. 34, 808–816.
- Long, J., *et al.*, 2025. Data-driven movement analysis. *International Journal of Geographical Information Science*, 39 (5), 945–950.
- Lou, X., *et al.*, 2025. Geoxcp: uncertainty quantification of spatial explanations in explainable ai. *International Journal of Geographical Information Science*, 1–31.
- Luca, M., *et al.*, 2021. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)*, 55 (1), 1–44.
- Mandal, A., Leavy, S., and Little, S., 2021. Dataset diversity: measuring and mitigating geographical bias in image search and retrieval. In: *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing*. 19–25.
- Mauro, G., *et al.*, 2022. Generating mobility networks with generative adversarial networks. *EPJ data science*, 11 (1), 58.
- McIntosh, J. and Yuan, M., 2005. Assessing similarity of geographic processes and events. *Transactions in GIS*, 9 (2), 223–245.
- Merchant, A., *et al.*, 2023. Scaling deep learning for materials discovery. *Nature*, 624 (7990), 80–85.
- Moran, P.A., 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37 (1/2), 17–23.
- Moreno-Torres, J.G., *et al.*, 2012. A unifying view on dataset shift in classification. *Pattern recognition*, 45 (1), 521–530.
- Moska, J., *et al.*, 2025. Obsr: Open benchmark for spatial representations. In: *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*. 670–681.
- Nijs, K.d., Omodei, E., and Sekara, V., 2025. Data bias in human mobility is a universal phenomenon but is highly location-specific. *arXiv preprint arXiv:2508.00149*.
- Nilforoshan, H., *et al.*, 2023. Human mobility networks reveal increased segregation in large cities. *Nature*, 624 (7992), 586–592.
- Noi, E., Rudolph, A., and Dodge, S., 2022. Assessing covid-induced changes in spatiotemporal structure of mobility in the united states in 2020: a multi-source analytical framework. *International Journal of Geographical Information Science*, 36 (3), 585–616.
- Nougnanke, B., Blanc, G., and Robert, T., 2025. How dataset diversity affects generalization in ml-based nids. In: *European Symposium on Research in Computer Security*. Springer, 269–288.
- Pan, S.J. and Yang, Q., 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22 (10), 1345–1359.
- Panaretos, V.M. and Zemel, Y., 2019. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6 (1), 405–431.
- Pappalardo, L., *et al.*, 2023. Future directions in human mobility science. *Nature computational science*, 3 (7), 588–600.
- Pourebrahim, N., *et al.*, 2019. Trip distribution modeling with twitter data. *Computers, Environment and Urban Systems*, 77, 101354.
- Robinson, C. and Dilkina, B., 2018. A machine learning approach to modeling human migration. In: *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–8.

- Rong, C., Ding, J., and Li, Y., 2024. An interdisciplinary survey on origin-destination flows modeling: Theory and techniques. *ACM Computing Surveys*, 57 (1), 1–49.
- Rong, C., *et al.*, 2025. A large-scale dataset and benchmark for commuting origin-destination flow generation. In: Y. Yue, A. Garg, N. Peng, F. Sha and R. Yu, eds. *International Conference on Representation Learning*. vol. 2025, 96180–96205.
- Rong, C., Feng, J., and Ding, J., 2023a. GODDAG: Generating origin-destination flow for new cities via domain adversarial training. *IEEE Transactions on Knowledge and Data Engineering*, 35 (10), 10048–10057.
- Rong, C., Wang, H., and Li, Y., 2023b. Origin-destination network generation via gravity-guided GAN. *arXiv preprint arXiv:2306.03390*.
- Santana, C., *et al.*, 2023. Covid-19 is linked to changes in the time–space dimension of human mobility. *Nature Human Behaviour*, 7 (10), 1729–1739.
- Schläpfer, M., *et al.*, 2021. The universal visitation law of human mobility. *Nature*, 593 (7860), 522–527.
- Schlosser, F., *et al.*, 2021. Biases in human mobility data impact epidemic modeling. *arXiv preprint arXiv:2112.12521*.
- Schwering, A., 2008. Approaches to semantic similarity measurement for geo-spatial data: a survey. *Transactions in GIS*, 12 (1), 5–29.
- Shi, M., *et al.*, 2025. Defining concept drift and its variants in research data management: A scientometric case study on geographic information science. *Transactions in GIS*, 29 (3), e70058.
- Simini, F., *et al.*, 2021. A deep gravity model for mobility flows generation. *Nature communications*, 12 (1), 6576.
- Stewart, A.J., *et al.*, 2025. Torchgeo: deep learning with geospatial data. *ACM Transactions on Spatial Algorithms and Systems*, 11 (4), 1–28.
- Sun, B., Feng, J., and Saenko, K., 2017. Correlation alignment for unsupervised domain adaptation. In: *Domain adaptation in computer vision applications*. Springer, 153–171.
- Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46 (sup1), 234–240.
- Wang, S. and Zhu, D., 2025. A deep origin-destination flow imputation model informed by the visitation law in human mobility. In: *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*. 1266–1269.
- Wang, S., *et al.*, 2024a. Infrequent activities predict economic outcomes in major american cities. *Nature Cities*, 1 (4), 305–314.
- Wang, Y., *et al.*, 2024b. Cola: Cross-city mobility transformer for human trajectory simulation. In: *Proceedings of the ACM Web Conference 2024*. 3509–3520.
- Wang, Z., *et al.*, 2025. Geobs: Information-theoretic quantification of geographic bias in ai models. *arXiv preprint arXiv:2509.23482*.
- Xu, A., 2023. Spatial patterns and determinants of inter-county migration in california: A multilevel gravity model approach. *Population Research and Policy Review*, 42 (3), 40.
- Xu, F., *et al.*, 2025a. Using human mobility data to quantify experienced urban inequalities. *Nature*

- Human Behaviour*, 1–11.
- Xu, Y., *et al.*, 2018. Human mobility and socioeconomic status: Analysis of singapore and boston. *Computers, Environment and Urban Systems*, 72, 51–67.
- Xu, Y., *et al.*, 2021. Tourism geography through the lens of time use: A computational framework using fine-grained mobile phone data. *Annals of the American Association of Geographers*, 111 (5), 1420–1444.
- Xu, Y., *et al.*, 2016. Another tale of two cities: Understanding human activity space using actively tracked cellphone location data. *Annals of the American Association of Geographers*, 106 (2), 489–502.
- Xu, Y., *et al.*, 2023. Urban dynamics through the lens of human mobility. *Nature Computational Science*, 3 (7), 611–620.
- Xu, Y., *et al.*, 2025b. Predicting human mobility flows in cities using deep learning on satellite imagery. *Nature Communications*, 16 (1), 10372.
- Yabe, T., *et al.*, 2025. Behaviour-based dependency networks between places shape urban economic resilience. *Nature human behaviour*, 9 (3), 496–506.
- Yan, H., 2024. Quantifying spatial similarity for use as constraints in map generalisation. *Journal of Spatial Science*, 69 (1), 23–42.
- Yang, J., *et al.*, 2026. Transferable human mobility network reconstruction with neurogravity. *Nature Computational Science*, 6 (6), 630–641.
- Yu, C., *et al.*, 2024. Harnessing llms for cross-city od flow prediction. In: *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*. 384–395.
- Yu, J., *et al.*, 2025. Exploring geo-transferability of deep neural network by developing comprehensive metrics. *Geo-spatial Information Science*, 1–19.
- Yuan, Y., *et al.*, 2025. Learning the complexity of urban mobility with deep generative network. *PNAS nexus*, 4 (5), pgaf081.
- Zhang, F., *et al.*, 2024. Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. *Annals of the American Association of Geographers*, 114 (5), 876–897.
- Zhang, X., *et al.*, 2026. City identity recognition: how representation bias influences model predictability and replicability? *Computers, Environment and Urban Systems*, 123, 102370.
- Zhao, D., *et al.*, 2024. Measuring diversity in datasets. In: *International Conference on Learning Representations*. vol. 1, 36.
- Zhao, F.H., *et al.*, 2025a. A multivariate spatial structure indicator based on geographic similarity. *International Journal of Geographical Information Science*, 39 (7), 1518–1539.
- Zhao, Y., *et al.*, 2025b. Predicting origin-destination flows by considering heterogeneous mobility patterns. *Sustainable Cities and Society*, 118, 106015.
- Zheng, Y., *et al.*, 2024. Impacts of remote work on vehicle miles traveled and transit ridership in the usa. *Nature Cities*, 1 (5), 346–358.
- Zhu, A.X., *et al.*, 2018. Spatial prediction based on third law of geography. *Annals of GIS*, 24 (4), 225–240.
- Zhu, A.X. and Turner, M., 2022. How is the third law of geography different? *Annals of GIS*, 28 (1),

57–67.

Zhu, A., *et al.*, 2015. Predictive soil mapping with limited sample data. *European Journal of Soil Science*, 66 (3), 535–547.

Zhu, D. and Ma, Z., 2026. Gravity-informed deep flow inference for spatial evolution modeling in panel data. *International Journal of Geographical Information Science*, 40 (4), 918–946.

Zhuang, F., *et al.*, 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109 (1), 43–76.